A rough association rule is applicable for knowledge discovery

Shu-Hsien Liao, Professor
Department of Management Sciences, Decision Making,
Tamkang University
Taipei, Taiwan
michael@mail.tku.edu.tw

Yin-Ju Chen, PhD Student
Department of Management Sciences, Tamkang University
Taipei, Taiwan
s5515124@ms18.hinet.net

***Abstract*—The traditional association rule which should be fixed in order to avoid both that only trivial rules are retained and also that interesting rules are not discarded. In fact, the situations which use the relative comparison to express are more complete than to use the absolute comparison. Through relative comparison we proposes a new approach for mining association rule, which has the ability to handle the uncertainty in the classing process, so that we can reduce information loss and enhance the result of data mining. In this paper, the new approach can be applied in find association rules, which has the ability to handle the uncertainty in the classing process and suitable for all data types.**

Keywords- Data mining, Association rule, Electronic commerce, Rough set

I. INTRODUCTION

Many algorithms have been proposed for mining Boolean association rules. However, very little work has been done m mining quantitative association rules. Although we can transform quantitative attributes into Boolean attributes, this approach Is not effective and is difficult to scale up for high-dimensional cases and also may result m many imprecise association rules [1].The application of association rules is not limited to marketing problems: in fact they can shed light on a wide range of knowledge discovery and decision making problems. The basic problem of mining association rules is then to generate all association rules A $\Rightarrow$ B that have support and confidence greater than user-specified thresholds [5].

The remainder of this paper is organized as follows. Section 2 the traditional Apriori algorithm. Section 3 the problem statement. Section 4 new algorithms modified from Apriori algorithm. Closing remarks and future work are presented in Sect. 5.

II. TRADITIONAL APRIORI ALGORITHM

Make $I = \{i_1, i_2, \cdots i_m\}$ the item set, in which each item represents a specific literal. D stands for a set of transactions in a database in which each transaction T represents an item set such that $T \subseteq I$. That is, each item set T is a non-empty sub-item set of I. The association rules are an implication of the form $X \rightarrow Y$, where $X \subset I, Y \subset I$ and $X \cap Y = \phi$. The rule $X \rightarrow Y$ holds in the transaction set D according to two measurement standards - support and confidence. Support (denoted as Sup(X,D)) represents the rate of transactions in D containing the item set X. Support is used to evaluate the statistical importance of D, and the higher its value, the more important the transaction set D is [4]. Fig.1 presents the process of traditional Apriorial inference.

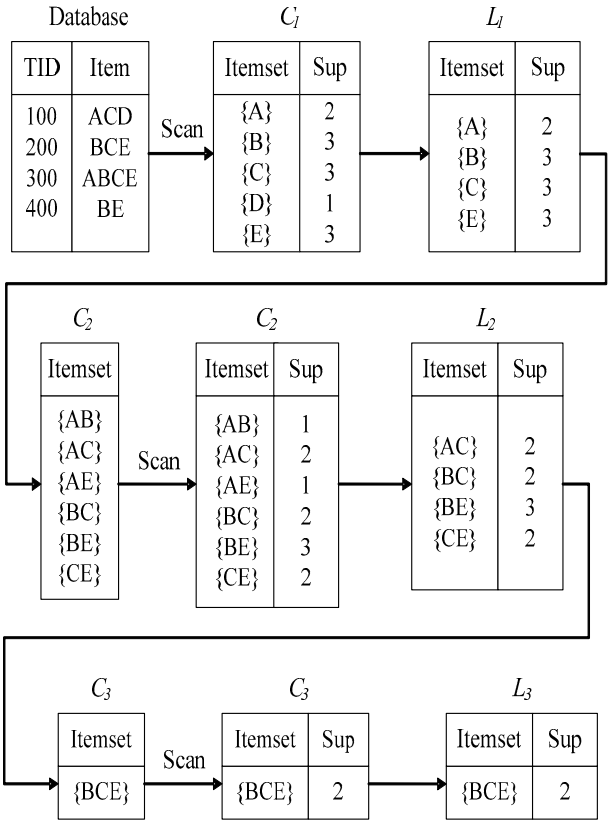

Figure1. The process of traditional Apriorial inference

III. PROBLEM STATEMENT

A. *From the angle of discussing Support and Confidence*

There are many different algorithms to discover association rules from various data types. First, Boolean attributes, which represent the items, are assumed to have only two values. If an item is in a transaction, the corresponding attribute value will be 1; otherwise, the value will be 0. By assuming that a variable may have more than two statuses, Boolean attributes can be generalized to nominal attributes. Many interesting and efficient algorithms have been proposed for mining association rules from Boolean or nominal data, such as Apriori and DHP.

978-1-4244-4738-1/09/$25.00 ©2009 IEEE

Second, since quantitative data are popular in practical databases, a natural extension is finding association rules from quantitative data. To solve this problem, previous research partitioned the value of a quantitative attribute into a set of intervals, so that the traditional algorithms for nominal data could be applied [2].

In mining association rules the two important measures are the support and the confidence. In generally, the mining of association rules is divided into two phases: finding frequent itemsets with support above a threshold, and finding rules with confidence above a threshold [3]. For example (table1), the minimum support is assumed to be30% and the minimum confidence is assumed to be 50%.

Table1: Sample data set

No	Attribute			Decision
	Age	Gender	Shopping frequency	Brand loyalty
t_1	20~29	Male	Once a month	High
t_2	30~39	Female	Under fortnightly	Low
t_3	20~29	Male	Once a month	High
t_4	20~29	Female	Once a month	Median
t_5	10~19	Male	other	Low
t_6	10~19	Male	fortnightly	Low
t_7	30~39	Female	Under fortnightly	Median
t_8	10~19	Male	other	Low
t_9	20~29	Male	Once a month	High
t_{10}	30~39	Female	Under fortnightly	Median

The support, confidence and lift value of the traditional association rule are defined as below.

$$\text{Support}\left(\left(\text{Male}\right)\cap\left(\text{Once a month}\right)\cap\left(\text{Age20}\sim 29\right)\right)=$$

$$\frac{\left(\text{Male}\right)\cap\left(\text{Once a month}\right)\cap\left(\text{Age20}\sim 29\right)\text{ Total of trades in database}}{\text{Total of trades in database}}=\frac{3}{10}$$

$$\text{Confidence}\left(\left(\text{Male}\right)\cap\left(\text{Once a month}\right)\cap\left(\text{Age20}\sim 29\right)\rightarrow\left(\text{Brand loyalty - high}\right)\right)=$$

$$\frac{\left(\text{Male}\right)\cap\left(\text{Once a month}\right)\cap\left(\text{Age20}\sim 29\right)\cap\left(\text{Brand loyalty - high}\right)\text{Total of trades in database}}{\left(\text{Male}\right)\cap\left(\text{Once a month}\right)\cap\left(\text{Age20}\sim 29\right)\text{ Total of trades in database}}=\frac{3}{3}$$

$$\text{Lift}=\frac{\text{Confidence}\left(\left(\text{Male}\right)\cap\left(\text{Once a month}\right)\cap\left(\text{Age20}\sim 29\right)\rightarrow\left(\text{Brand loyalty}\right)\right)}{\text{Support}\left(\text{Brand loyalty - high}\right)}=\frac{3/3}{3/10}=\frac{10}{3}>1$$

Rules express the relation between pairs of items and are defined in two measures: support and confidence. Most techniques for finding association rule scan the whole data set, evaluate all possible rules and retain only rules that have support and confidence greater than thresholds. It's mean that the situations which use the absolute comparison. Actually, it's not reasonable. For instance, if we raise the threshold of support to 50%, we can not find the association rule—a male who is 20~29 and go to shopping once a month has high brand loyalty. The traditional association rule which should be fixed in order to avoid both that only trivial rules are retained and also that interesting rules are not discarded. In fact, the situations which use the relative comparison to express are more complete than to use the absolute comparison. Through relative comparison we proposes a new approach for mining association rule, which has the ability to handle the uncertainty in the classing process, so that we can reduce information loss and enhance the result of data mining.

The remaining transactions can still be partitioned into those that actually violate the rule, and those which do not carry any relevant information [5].

B. From the angle of information disclosure

The ordinal data sets and Boolean data types are as table2 and table3. First, if we want to know a male who is 20~25 and go to purchasing Coca-cola. The one situation is the male just buying Coca-cola, and the other situation is the male who is buying both Coca-cola and Pepsi.

Using the traditional association rule for data mining, first to count value of $\text{Support}(\text{male}\cap 20\sim 25\cap \text{Coca}-\text{cola})$. Then to count value of $\text{Support}(\text{male}\cap 20\sim 25\cap \text{Pepsi})$. Finally, the same customer is the one who is buying both Coca-cola and Pepsi

Table2: ordinal data set

No	Gender	Age	Purchasing product
001	Male	20	Coca-cola & Pepsi
002	Female	23	Pepsi
003	Female	17	Pepsi
004	Male	30	Coca-cola
005	Male	22	Coca-cola & Pepsi

Table3: Data types (Boolean attributes)

No	Gender		Age			Purchasing product	
	Male	Female	Under 20	20~25	26~30	Coca-cola	Pepsi
001	1	0	0	1	0	1	1
002	0	1	0	1	0	0	1
003	0	1	1	0	0	0	1
004	1	0	0	0	1	1	0
005	1	0	0	1	0	1	1

According to the basic concepts of support and confidence, there are four situations to mining association rule.

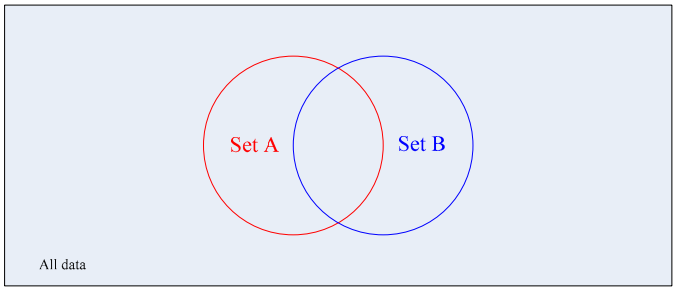

Figure2. Situation I (Set A and Set B are both small)

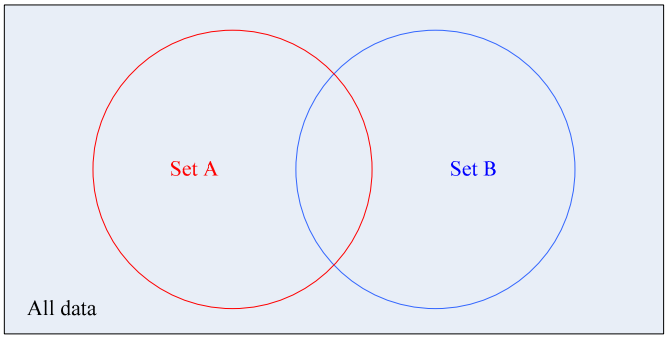

Figure3. Situation II (Set A and Set B are both big)

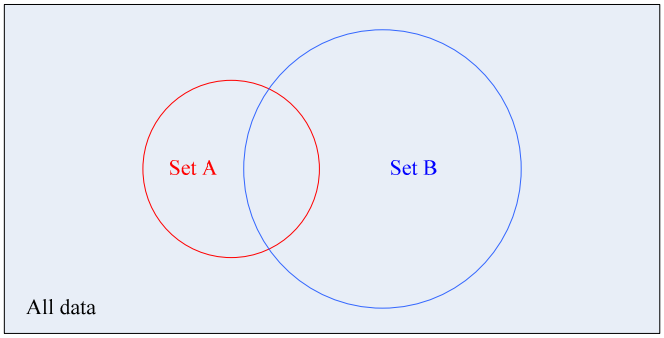

Figure4. Situation III (Set B is bigger than Set A)

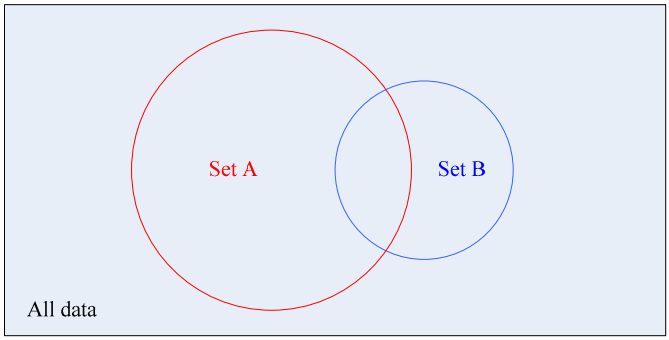

Figure5. Situation IV (Set A is bigger than Set B)

The technical terminologies—support and confidence value of the traditional association rule, as show in fig.6. to fig.13.

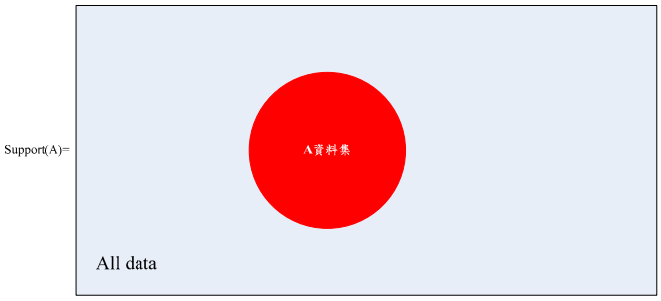

Figure6. Support value of the traditional association rule in situation I

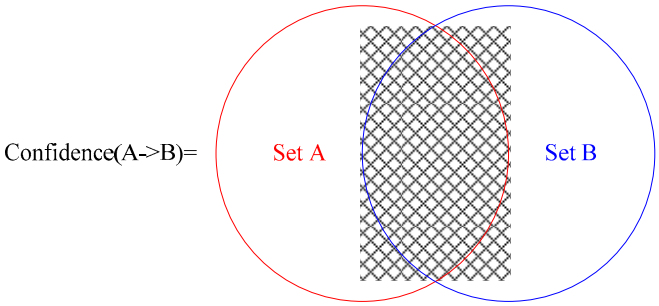

Figure7. Confidence value of the traditional association rule in situation I

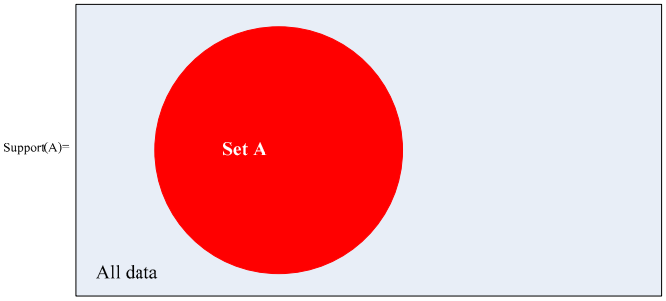

Figure8. Support value of the traditional association rule in situation II

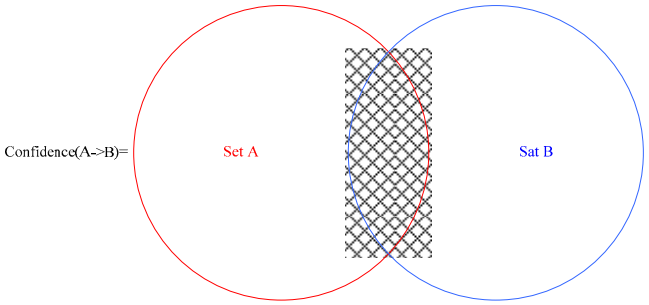

Figure9. Confidence value of the traditional association rule in situation II

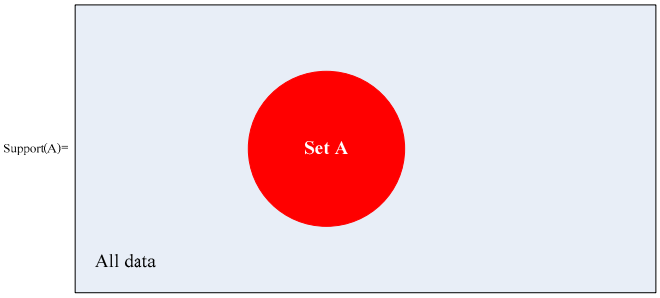

Figure10. Support value of the traditional association rule in situation III

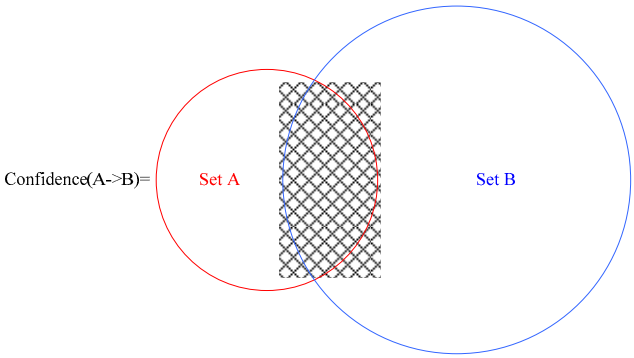

Figure11. Confidence value of the traditional association rule in situation III

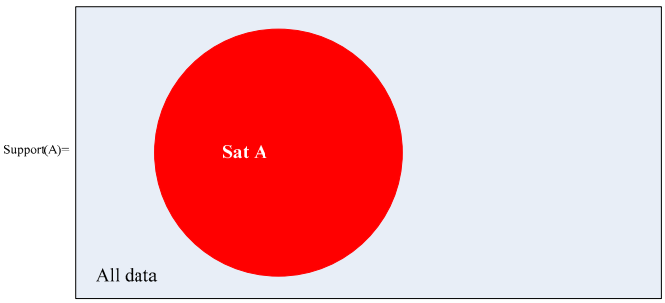

Figure12. Support value of the traditional association rule in situation IV

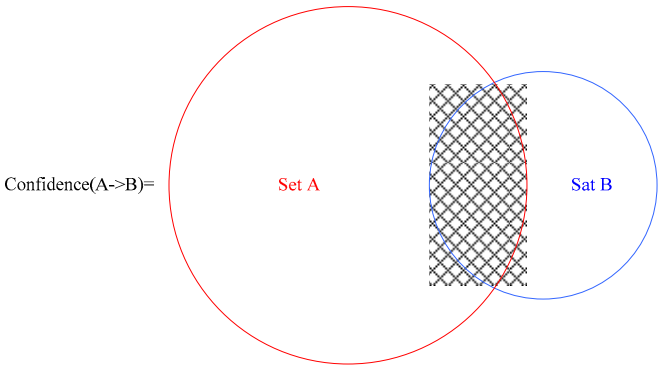

Figure13. Confidence value of the traditional association rule in situation IV

IV. METHODOLOGY

As fig.4, fig.5 and fig.10 to fig.13 is not applicable for tradition association rules of mining that have support and confidence greater than user-specified threshold. For example, the sample data set just has set A and set B. Set A occupies most of the sample date set (70%), and the remainder is set B. If we determining the degree of min support of 40%, we can not find the association rule—about set B. If we decrease the threshold of min support to 2%, we get too much association rules—about set B.

In other words, the excessively low min support will generate too many association rules. And the excessively high min support will bring too few association rules. Many decision makers are persecuted by using tradition rules when determining the degree of support.

In order to solve the kind of problem, the aim of the research is to provide a new association rule concept, which is using Bayesian network. The new association rule algorithm is show as below.

$$\text{Support}(X) = P(\frac{X}{T_i}) = P(X|T_i) = P(\frac{X \cap T_i}{T_i});\ X \subseteq T \quad (1)$$

$$\text{Conf}(X \to Y) = P(\frac{X \cap Y}{X}) = P(X \cap Y|X) \quad (2)$$

$P(X) = P(X|T_i)$: prior probability

$P(X \cap Y|X)$: sample probability

$P(X|X \cap Y)$: posterior probability

$$P(X|X \cap Y) = \frac{P(X \cap Y|X)P(X)}{P(X \cap Y)} = \frac{P(X \cap Y|X)P(X)}{\sum P(X \cap Y|X)P(X)} \quad (3)$$

Through Bayesian theory, we get prior probability which include both prior probability and sample probability. From the angle of information disclosure, compares with the traditional association rules, decision maker can get more information from Bayesian association rules.

When we get marketing survey, how to use the algorithm is as below, which include both rough set theory and Bayesian theory.

A. *Step1:Data Processing*

Each answer regards which from questionnaire fills are as a rule.

X_i^c : answer regards

c: answer's number from 1 to m

i: questionnaire's number from 1to n

A_i^j : questionnaire fills; $j = 1 \cdots o$

As table1, all answers can be defined as

$$\mu_{c_j}^{(\underline{x})} = \bigcap_{i=1}^{n} \mu_{A_{ij}}(x_i),\ \underline{x} = [x_1, x_2 \cdots x_n]^T \quad (4)$$

B. *Step2: To sieve out the same rules*

In order to sieve out the same rules, all answers can be defined as

$$\mu_{R_j}(x_i) = \{\mu_{c_j}(\underline{x})\} = \bigcap_{i=1}^{n} \mu_{A_j}(x_i) \quad (5)$$

C. *Step3: Determined the core attributes value*

It embraces the core attributes value concept of rough set.

$$\text{lower}\mu_{Rj} = \{\mu_{Rj}(x_i) \in \mu_{Aij}(x_i) \left| [\mu_{Rj}(x_i)]_{\text{core}} \subset \mu_{cj}(\underline{x}) \right.\} \quad (6)$$

$$\text{upper}\mu_{Rj} = \{\mu_{cj}(\underline{x}) \in \mu_{Aij}(x_i) \left| [\mu_{Rj}(x_i)]_{\text{core}} \cap \mu_{cj}(\underline{x}) \neq 0 \right.\} \quad (7)$$

The core attributes value concept of rough set is show as fig.14.

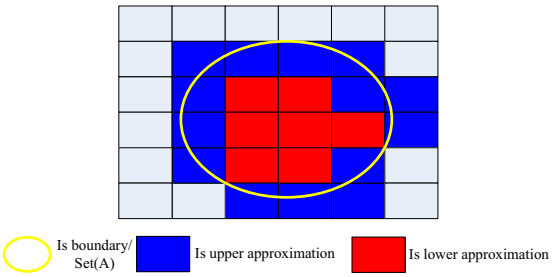

Figure14.

D. Step4: Compute Reliability

The reliability is defined as

$$R_B = \frac{\text{Lower}}{\text{Upper}} \tag{8}$$

E. Step5: Application of the posterior probability Inference concept to finding Bayesian association rule

$$\text{Support}(X) = P(\frac{X}{T_i}) = P(X|T_i) = P(\frac{X \cap T_i}{T_i});\ X \subseteq T \tag{9}$$

$$\text{Conf}(X \to Y) = P(\frac{X \cap Y}{X}) = P(X \cap Y|X) \tag{10}$$

$$P(X|X \cap Y) = \frac{P(X \cap Y|X)P(X)}{P(X \cap Y)} = \frac{P(X \cap Y|X)P(X)}{\sum P(X \cap Y|X)P(X)} \tag{11}$$

V. CONCLUSION

A. The benefit of the new association rule algorithm which include rough set theory

Using the approximate concept of rough set theory, decision maker will get more information than tradition association rules. The user dose not has to determining the degree of support and confidence for thresholds. The situations which use the relative comparison to express are more complete than to use the absolute comparison. The new approach for mining association rule, which has the ability to handle the uncertainty in the classing process, so that we can reduce information loss and enhance the result of data mining.

B. The benefit of the new association rule algorithm which include Bayesian network

Using the new algorithm can simulate the value of probability, which is based on the continuous data set. According to the correlation between random variable to displays group's between correlations. And it also can be used of solving the mistake which is induced by the insufficient knowledge of user and the mistakes made by input error.

ACKNOWLEDGEMENTS

This research was funded by the National Science Council, Taiwan, Republic of China, under contract No. NSC 98-2410-H-032 -038 – MY2.

REFERENCES

[1] Lian, Wang; Cheung, David W.; Yiu, S.M.," An efficient algorithm for finding dense regions for mining quantitative association rules", Computers and Mathematics with Applications, Vol.50, No.3-4, 2005, pp. 471-490.

[2] Chen, Yen-Liang; Weng, Cheng-Hsiung, "Mining association rules from imprecise ordinal data", Fuzzy Sets and Systems, Vol. 159, 2008, pp. 460-474.

[3] Boris Rozenberg, Ehud Gudes, "Association rules mining in vertically partitioned database", Data & Knowledge Engineering, Vol.59, 2006, pp. 378-396.

[4] S. H. Liao, C. M. Chen., and .C. H. Wu, (2008) ₃ Mining customer knowledge for product line and brand extension in retailing₄ , Expert Systems with Applications, Vol. 35, Issue. 3. pp. 1763-1776.

[5] De Cock, M.; Cornelis, C.; Kerre, E.E.,(2005)" Elicitation of fuzzy association rules from positive and negative examples" Fuzzy Sets and Systems, Vol. 149, Issue. 1. pp. 73-85